



**KB en migratie**  
*Werkdocument*

Uitgave: 0.2  
Auteur: Caroline van Wijk  
Datum: 7 augustus 2006

Koninklijke Bibliotheek  
Afdeling Digitale Duurzaamheid

## KB en migratie

---

### Ontstaansgeschiedenis KB en migratie

Versie Document	Datum wijziging	Auteur	Samenvatting wijzigingen
0.1	4 juli 2006	Caroline van Wijk	
0.2	7 augustus 2006	Caroline van Wijk	Opmerkingen H. van Wijngaarden en J. Rog verwerkt. <ul style="list-style-type: none"><li>• Toegevoegd: karakterisering van een digitaal object op een functionele én technische karakterisering toegevoegd.</li><li>• Uitgebreid: normalisatie is een optie voor elk aanbod van niet-gangbare bestandsformaten, de bestanden van DARE zijn een concreet voorbeeld hiervan.</li><li>• Paragraaf 2.3 Karakterisering digitale objecten aangepast</li></ul>

### Aanverwante documenten

*Project Initiatie Document – Migratieonderzoek, C. van Wijk (2006)*

*Startpunt Migratieonderzoek, C. van Wijk (2006)*

*Testplan Migratieonderzoek, C. van wijk (2006)*

## Inhoudsopgave

<b>1</b>	<b>Inleiding .....</b>	<b>4</b>
1.1	Doel van dit document	4
1.2	Indeling van dit document	4
	<i>Inhoud e-Depot .....</i>	<i>4</i>
	<i>Mogelijkheden en consequenties van migratie voor de KB .....</i>	<i>4</i>
	<i>Bijlage inhoud e-Depot .....</i>	<i>5</i>
<b>2</b>	<b>Inhoud e-Depot .....</b>	<b>6</b>
2.1	Inleiding	6
2.2	Gearchiveerde objecten van het e-Depot	6
2.3	Karakterisering digitale objecten	6
<b>3</b>	<b>Mogelijkheden en consequenties van migratie voor de KB.....</b>	<b>8</b>
3.1	Inleiding	8
3.2	Bestandsformatenonderzoek	8
3.3	Test Migratieonderzoek	8
3.4	Toepassing migratie voor de KB	9
<b>I</b>	<b>Bijlage inhoud e-Depot .....</b>	<b>10</b>
I.1	Voorbeeld functionele karakterisering van een e-Depot artikel .....	11
	<i>Inhoud .....</i>	<i>12</i>
	<i>Structuur .....</i>	<i>14</i>
	<i>Uiterlijk.....</i>	<i>16</i>
	<i>Context .....</i>	<i>20</i>
	<i>Gedrag .....</i>	<i>21</i>
1.2	Voorbeeld technische karakterisering van een e-Depot artikel .....	22
1.3	Overzicht (voorlopige) functionele karakterisering van inhoud e-Depot.....	25

## 1 Inleiding

### 1.1 Doel van dit document

In dit werkdokument worden de huidige mogelijkheden (zomer 2006) wat betreft migratie als digitale duurzaamheidsstrategie voor de Koninklijke Bibliotheek beschreven.

Dit document bestaat uit een overzicht van de toepasbaarheid van verschillende werkwijzen van migratie voor de KB (zie *Startpunt Migratieonderzoek*). De toepasbaarheid van migratie voor de KB is gebaseerd op de (toekomstige) inhoud van het e-Depot en de resultaten van het Bestandsformatenonderzoek wat betreft bestandsformaatveroudering. Het Bestandsformatenonderzoek wordt op het moment van schrijven van dit document nog uitgevoerd. Eventuele resultaten uit het Bestandsformatenonderzoek zullen alsnog verwerkt worden.

Dit document wordt als startpunt voor het testplan voor migratietools gebruikt.

Opmerkingen en toevoegingen zijn welkom en kunnen naar de auteur verstuurd worden ([caroline.vanwijk@kb.nl](mailto:caroline.vanwijk@kb.nl)).

### 1.2 Indeling van dit document

#### Inhoud e-Depot

In dit hoofdstuk wordt beschreven welke typen objecten in het e-Depot zijn opgeslagen en zullen worden opgeslagen.

Ook wordt een functionele en technische karakterisering van het meest voorkomende type object gegeven. Deze karakterisering kan dienen als raamwerk voor het testen van conversie van e-Depot objecten. Voor het bepalen van het resultaat van een conversie/migratie (gelukt of niet) is het nodig om de essentie van een digitaal object te bepalen. Functioneel gezien kan dit aan de hand van de vijf eigenschappen van een digitaal object: inhoud, structuur, uiterlijk, gedrag en context. Een technische karakterisering hangt samen met het bestandsformaat.

#### Mogelijkheden en consequenties van migratie voor de KB

In dit deel van het document worden mogelijkheden en consequenties van het gebruik van migratie beschreven op basis van de inhoud van het e-Depot en de op dit moment bekende migratie werkwijzen.

Resultaten uit het Bestandsformatenonderzoek, die van belang zijn voor migratie, worden besproken.

### **Bijlage inhoud e-Depot**

In de bijlage wordt een voorbeeld gegeven van de functionele en technische karakterisering van een wetenschappelijk artikel uit het e-Depot. Ook wordt een eerste opzet voor een functionele indeling van de objecten in het e-Depot beschreven.

## 2 Inhoud e-Depot

### 2.1 Inleiding

In dit hoofdstuk worden gearchiveerde objecten in het e-Depot en toekomstige objecten per bestandsformaat in een lijst geplaatst.

Op basis van de resultaten van het Bestandsformatenonderzoek wordt aangegeven welke type objecten “in gevaar zijn”.

De karakterisering van het in het e-Depot meest voorkomend type object wordt aan het eind van dit hoofdstuk beschreven.

### 2.2 Gearchiveerde objecten van het e-Depot

Per februari (?) 2006

PDF 1.1	288.823
PDF 1.2	1.623.991
PDF 1.3	1.902.242
PDF 1.4	28.422
PDF 1.5	15.449
TIFF 5.0	4.163
TIFF 6.0	273.281
JPG	252.107 (niet als hoofdbestand aangegeven, dus als bijlage)

Het DARE project zal een heterogeen aanbod van publicaties toevoegen aan de inhoud van het e-Depot:

- MS Office (Worddocumenten)
- Photoshop-documenten.
- Etc.

Het webarchivering project zal websites als publicaties toevoegen aan de inhoud van het e-Depot:

- Websites kunnen bestaan uit platte html, verschillende soorten beeldformaten (JPEG, GIF etc.) Flash, JavaScript, andere vormen van scripting, ASP, databases, etc.

### 2.3 Karakterisering digitale objecten

Voor het opstellen van migratie trajecten en procedures is het nodig om te bepalen wat de eigenschappen van het originele bestand zijn en welke eigenschappen ook in het doelbestand aanwezig moeten zijn.

Digitale objecten karakteriseren op basis van bestandsformaat wordt met de uitbreiding van veel formaten steeds moeilijker. Een bestandsformaat als PDF krijgt steeds meer

## KB en migratie

---

functionaliteit en is niet meer standaard als “tekstdocument” aan te merken als gebruik wordt gemaakt van nieuwe functionaliteit zoals het opnemen van filmpjes in een PDF.

Het is nodig, om naast een indeling op bestandsformaat, ook een functionele indeling te maken. Een functionele indeling is gebaseerd op de functie die het digitale object heeft voor het publiek. Een functionele indeling zou als categorieën “een publicatie uit een wetenschappelijk tijdschrift” of een “informatieve website van een instelling” kunnen hebben.

Op dit moment zijn er niet veel opties voor automatische karakterisering. Het programma Jhove genereert technische metadata (MIX elementen) zoals welke fonts embedded of gesubset zijn en hoe groot het bestand is. Dit is karakterisering op bestandsniveau. Voor migratie procedures is een indeling wat betreft de karakteristieken op een hoger niveau dan bestandsniveau nodig. Er zijn teveel objecten in het e-Depot opgeslagen om per opgeslagen object of per bestand een migratieprocedure te bepalen.

Functionele karakterisering wat betreft de 5 eigenschappen van een digitaal object (inhoud, context, uiterlijk, gedrag en structuur) is op dit moment alleen handmatig te bepalen.

De migratieprocedures van dit onderzoek zijn gebaseerd op migratie van een publicatie uit een wetenschappelijk tijdschrift en de bijbehorende karakterisering.

In de bijlage van dit document is een voorbeeld van karakterisering van een wetenschappelijke publicatie opgenomen. Publicatie uit een wetenschappelijk tijdschrift is op dit moment de hoofdcategorie van de gearchiveerde objecten in het e-Depot. Een eerste opzet van een functionele indeling van de inhoud van het e-Depot is ook toegevoegd aan het document.

### 3 Mogelijkheden en consequenties van migratie voor de KB

#### 3.1 Inleiding

In dit hoofdstuk worden de mogelijkheden en consequenties van migratie voor de KB besproken aan de hand van onder meer voorlopige resultaten uit het Bestandsformatenonderzoek. Het Bestandsformatenonderzoek bekijkt welke bestandsformaten, die in het e-Depot voorkomen, met ontoegankelijkheid bedreigd worden en eventueel in aanmerking komen voor migratie.

#### 3.2 Bestandsformatenonderzoek

Onderzoek naar eventueel verlies van toegankelijkheid van de in het e-Depot opgeslagen publicaties is nog gaande (Bestandsformaten onderzoek). Vooral nog lijkt geen versie van PDF of TIFF bedreigd met ontoegankelijkheid, maar een gerichte test met behulp van een steekproef moet nog uitgevoerd worden (Bestandsformatenonderzoek). Het overzicht van mogelijkheden en consequenties, dat hier besproken wordt, is dus nog niet volledig. Een lijst met bestandsformaten uit het e-Depot die bedreigd worden met bestandsveroudering is nog niet beschikbaar.

[resultaten Bestandsformatenonderzoek]

#### 3.3 Test Migratieonderzoek

Het is tóch interessant om een migratietest uit te voeren met de in het e-Depot opgeslagen digitale publicaties, ook al zijn deze digitale objecten nog niet bedreigd met ontoegankelijkheid.

Aan de hand van een test en het resultaat kunnen volgende onderdelen opgezet worden:

- Ontwikkelen van test procedure
- Ontwikkelen van migratie procedures
- In kaart brengen van benodigde tools
- In kaart brengen van openstaande onderdelen, wat betreft migratieprocessen, die nog niet zijn getest
- Informatie verzamelen die gebruikt kan worden voor het opzetten van een registry voor preservation tools (een registry is onderdeel van het project PLANETS)

### 3.4 Toepassing migratie voor de KB

Opties voor migratie in de huidige situatie:

- Normalisatie van een aanbod aan niet-gangbare bestandsformaten

Het DARE project is een concreet voorbeeld hiervan en zal een heterogeen aanbod aan bestandsformaten hebben vanuit de Nederlandse universiteiten. Waarschijnlijk zullen ook niet-gangbare formaten worden aangeleverd. Voor het DARE project kan normalisatie van bestandsformaten een optie zijn.

Consequenties:

- Het opslaan van een genormaliseerde versie én het origineel neemt ongeveer twee maal zoveel ruimte in het e-Depot in beslag.
- De genormaliseerde versie heeft ook een digitale duurzaamheid strategie nodig. Het hebben van een digitale duurzaamheidsstrategie voor een paar standaardformaten kost echter minder inspanning dan voor een hele reeks formaten.

Opties voor migratie in een toekomstige situatie:

- Migratie naar nieuwere versie van eenzelfde bestandsformaat

Consequenties:

- Het Nationaal Archief heeft met het project Testbed tests gedaan met stapsgewijze migratie naar nieuwere versies van eenzelfde bestandsformaat. Zij hebben dit vergeleken met migratie naar nieuwere versies waarbij enkele versies werden overgeslagen. De uitkomst was dat stapsgewijs migreren meer fouten had gegenereerd dan sprongsgewijs migreren.

- Migratie op verzoek

Het CAMiLEON project en LOCKSS hebben een test gedaan met “migration on request”. Bij beide projecten is er sprake van een positief resultaat: migratie op verzoek is uitvoerbaar. En als er goed nagedacht wordt over de conversietool is deze voor een gedeelte ook te gebruiken voor de lange termijn.

- UVC

Een UVC voor JPG en GIF is al ontwikkeld. Een UVC voor PDF is een te complex project om binnen het migratieonderzoek onder te brengen.

## I Bijlage inhoud e-Depot

## I.1 Voorbeeld functionele karakterisering van een e-Depot artikel

Voorbeeldartikel is een PDF versie 1.3 bestand, getoond in Adobe Acrobat Reader 7.0.0. 14-12-2004

Artikel:

### **Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays**

**Shigeyuki Matsui<sup>1,2</sup>** ✉

<sup>1</sup>Department of Pharmacoepidemiology, School of Public Health, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan

<sup>2</sup>Translational Research Informatics Center, Foundation for Biomedical Research and Innovation, Minatojima-minami-machi, Chuo-ku, Kobe 650-0047, Japan

*BMC Bioinformatics* 2006, **7**:156 doi:10.1186/1471-2105-7-156

Met een blauwe **markering** wordt aangegeven waar eventuele “problemen” bij het lezen en begrijpen van het e-Depot artikel zouden kunnen optreden.

Ik heb bij het verdelen van het artikel over de vijf eigenschappen *inhoud*, *structuur*, *uiterlijk* (*appearance*), *context* en *gedrag* geprobeerd elke eigenschap los van de anderen te behandelen. Bijv. bij *inhoud* speelt de grootte van de letter geen rol. Maar bij *structuur* speelt de onderverdeling in kop en subkop wel een rol, echter het is *uiterlijk* dat bepaald dat een kop en subkop door een grotere en kleinere letter worden aangegeven. Per voorbeeld van de eigenschappen is een deel van het artikel gebruikt.

Voor het karakteriseren van digitale objecten is een onderverdeling op file formaat alleen niet aan te bevelen, nu vele formaten functioneel worden uitgebreid. Een indeling op type bestand naast een indeling op bestandsformaat is nuttig. Onder “type bestand of object” wordt een indeling als tekstdocumenten, afbeeldingen (statisch), film en spreadsheets verstaan. Een verdere uitwerking (in de richting van het functionele doel van het digitale object) van deze onderverdeling kan via een matrix of een boomstructuur toegepast worden. Hieronder een voorbeeld:

Tekstdocumenten	Afbeeldingen
*Artikel	*...
*Handleiding	
*Documentatie	

of

Type object	Functioneel doel van het object		
	Artikel	Handleiding	Documentatie
Tekstdocumenten			
Afbeeldingen			
Spreadsheets			

## Inhoud

### Background

Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive, gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies for various cancers including diffuse large B-cell lymphoma [1], follicular lymphoma [2], acute myeloid leukemia [3], lung adenocarcinoma [4], and metastatic kidney cancer [5]. The selected genes will be further investigated in subsequent studies using technically simpler, but more reliable assays such as multiplexed quantitative reversetranscriptase polymerase chain reaction (RT-PCR) in formalin-fixed, paraffin-embedded tissue sections for routine clinical use [6,7]. Accordingly, the primary task in

Published: 20 March 2006

*BMC Bioinformatics* 2006, **7**:156 doi:10.1186/1471-2105-7-156

Received: 01 September 2005

Accepted: 20 March 2006

This article is available from: <http://www.biomedcentral.com/1471-2105/7/156>

© 2006 Matsui; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*BMC Bioinformatics* 2006, **7**:156 <http://www.biomedcentral.com/1471-2105/7/156>

Page 2 of 9

*(page number not for citation purposes)*

early-phased prognostic marker studies with microarrays would be to select a small fraction of relevant genes for subsequent studies. To this end, multiple testing to identify genes associated with prognosis is typically adopted as primary analysis, which may provide a list of significant genes.

Prediction analysis using subsets of significant genes may supplement the primary analysis. It can provide information regarding predictive capability for subsets of significant genes. More importantly, provided that appropriate measures of predictive accuracy for survival outcomes are established, it may indicate another 'cut-off' for a list of significant genes on the basis of predictive accuracy through *gene filtering* other than the criteria to control false positives in multiple testing. Despite many methods for prediction analysis of survival outcomes proposed in bioinformatics and biostatistics literature, including application of partial least squares [8-10] and ridge regression

[11,12], most methods intend to use the full set of genes for prediction without regard to the primary analysis. In this article, we develop a methodology for predicting survival outcomes using subsets of significant genes. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. Here, given the first two components, we can perform gene filtering. For each component, we consider particular specifications or procedures to illustrate the methodology. In building prediction models, we assume Cox proportional hazard models with a compound covariate [13,14]. In assessing performance of prediction models, measures of explained variation for Cox regression models [15-17] may not aim to measure the performance of prediction models on future patients, i.e., predictive accuracy. We propose to use the cross-validated log partial likelihood [18,19] to measure predictive accuracy. To assess significance of prediction results, we apply the permutation procedures in cross-validated prediction proposed by Radmacher et al. [14]. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors, because it is important to determine whether a new genetic marker adds prognostic information to that already contained in the more established prognostic factors [20]. The performance of the methodology will be evaluated using simulated data and real data from a lymphoma study.

# KB en migratie

---

## Structuur

### *Pagina 1*

#### *Linkerkolom*

Paragraafkop: Background

Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive, gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies for various cancers including diffuse large B-cell lym-

#### *Rechterkolom*

phoma [1], follicular lymphoma [2], acute myeloid leukemia [3], lung adenocarcinoma [4], and metastatic kidney cancer [5]. The selected genes will be further investigated in subsequent studies using technically simpler, but more reliable assays such as multiplexed quantitative reversetranscriptase polymerase chain reaction (RT-PCR) in formalin-fixed, paraffin-embedded tissue sections for routine clinical use [6,7]. Accordingly, the primary task in

### *Pagina 2*

#### *Linkerkolom*

*BMC Bioinformatics* 2006, **7**:156 <http://www.biomedcentral.com/1471-2105/7/156>

Page 2 of 9

*(page number not for citation purposes)*

early-phased prognostic marker studies with microarrays would be to select a small fraction of relevant genes for subsequent studies. To this end, multiple testing to identify genes associated with prognosis is typically adopted as primary analysis, which may provide a list of significant genes. Prediction analysis using subsets of significant genes may supplement the primary analysis. It can provide information regarding predictive capability for subsets of significant genes. More importantly, provided that appropriate measures of predictive accuracy for survival outcomes are established, it may indicate another 'cut-off' for a list of

significant genes on the basis of predictive accuracy through *gene filtering* other than the criteria to control false positives in multiple testing. Despite many methods for prediction analysis of survival outcomes proposed in bioinformatics and biostatistics literature, including application of partial least squares [8-10] and ridge regression [11,12], most methods intend to use the full set of genes for prediction without regard to the primary analysis. In this article, we develop a methodology for predicting survival outcomes using subsets of significant genes. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. Here, given the first two components, we can perform gene filtering. For each component, we consider particular specifications or procedures to illustrate the methodology. In building prediction models, we assume Cox proportional hazard models with a compound covariate [13,14]. In assessing performance of prediction models, measures of explained variation for Cox regression models [15-17] may not aim to measure the performance of prediction models on future patients, i.e., predictive accuracy. We propose to use the cross-validated log partial likelihood [18,19] to measure predictive accuracy. To assess significance of prediction results, we apply the permutation procedures in cross-validated prediction proposed by Radmacher et al. [14]. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors, because it is important to determine whether a new genetic marker adds prognostic information to that already contained in the more established prognostic factors [20]. The performance of the methodology will be evaluated using simulated data and real data from a lymphoma study.

**Uiterlijk**

## Background

Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive, gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies for various cancers including diffuse large B-cell lym-

phoma [1], follicular lymphoma [2], acute myeloid leukemia [3], lung adenocarcinoma [4], and metastatic kidney cancer [5]. The selected genes will be further investigated in subsequent studies using technically simpler, but more reliable assays such as multiplexed quantitative reverse-transcriptase polymerase chain reaction (RT-PCR) in formalin-fixed, paraffin-embedded tissue sections for routine clinical use [6,7]. Accordingly, the primary task in

Methodology article

Open Access

## Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays

Shigeyuki Matsui\*<sup>1,2</sup>

Address: <sup>1</sup>Department of Pharmacoepidemiology, School of Public Health, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan and <sup>2</sup>Translational Research Informatics Center, Foundation for Biomedical Research and Innovation, Minatojima-minami-machi, Chuo-ku, Kobe 650-0047, Japan

Email: Shigeyuki Matsui\* - matsui@pbh.med.kyoto-u.ac.jp

\* Corresponding author

Published: 20 March 2006

Received: 01 September 2005

BMC Bioinformatics 2006, 7:156 doi:10.1186/1471-2105-7-156

Accepted: 20 March 2006

This article is available from: <http://www.biomedcentral.com/1471-2105/7/156>

© 2006 Matsui; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies. The primary analytical task in such studies is to select a small fraction of relevant genes, typically from a list of significant genes, for further investigation in subsequent studies.

**Results:** We develop a methodology for predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. As particular specifications, we assume Cox proportional hazard models with a compound covariate. For assessing predictive accuracy, we propose to use the cross-validated log partial likelihood. To assess significance of prediction results, we apply permutation procedures in cross-validated prediction. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors. The methodology is evaluated using both simulated and real data.

**Conclusion:** The developed methodology for prognostic prediction using a subset of significant genes can provide new insights based on predictive capability, possibly incorporating standard prognostic factors, in selecting a fraction of relevant genes for subsequent studies.

### Background

Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive, gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies for various cancers including diffuse large B-cell lym-

phoma [1], follicular lymphoma [2], acute myeloid leukemia [3], lung adenocarcinoma [4], and metastatic kidney cancer [5]. The selected genes will be further investigated in subsequent studies using technically simpler, but more reliable assays such as multiplexed quantitative reverse-transcriptase polymerase chain reaction (RT-PCR) in formalin-fixed, paraffin-embedded tissue sections for routine clinical use [6,7]. Accordingly, the primary task in

early-phased prognostic marker studies with microarrays would be to select a small fraction of relevant genes for subsequent studies. To this end, multiple testing to identify genes associated with prognosis is typically adopted as primary analysis, which may provide a list of significant genes.

Prediction analysis using subsets of significant genes may supplement the primary analysis. It can provide information regarding predictive capability for subsets of significant genes. More importantly, provided that appropriate measures of predictive accuracy for survival outcomes are established, it may indicate another 'cut-off' for a list of significant genes on the basis of predictive accuracy through *gene filtering* other than the criteria to control false positives in multiple testing. Despite many methods for prediction analysis of survival outcomes proposed in bioinformatics and biostatistics literature, including application of partial least squares [8-10] and ridge regression [11,12], most methods intend to use the full set of genes for prediction without regard to the primary analysis.

In this article, we develop a methodology for predicting survival outcomes using subsets of significant genes. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. Here, given the first two components, we can perform gene filtering. For each component, we consider particular specifications or procedures to illustrate the methodology. In building prediction models, we assume Cox proportional hazard models with a compound covariate [13,14]. In assessing performance of prediction models, measures of explained variation for Cox regression models [15-17] may not aim to measure the performance of prediction models on future patients, i.e., predictive accuracy. We propose to use the cross-validated log partial likelihood [18,19] to measure predictive accuracy. To assess significance of prediction results, we apply the permutation procedures in cross-validated prediction proposed by Radmacher et al. [14]. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors, because it is important to determine whether a new genetic marker adds prognostic information to that already contained in the more established prognostic factors [20]. The performance of the methodology will be evaluated using simulated data and real data from a lymphoma study.

**Results**

**Gene filtering**

The simplest approach of gene filtering is based on the marginal association between each gene expression and

survival time [1-5]. For patient *i* in the training set, let  $h_i(t)$  be the hazard function and  $x_{ji}$  be the expression level for gene *j*. For gene *j*, we assume the univariate Cox regression model,

$$h_i(t) = h_{j,0}(t) \exp(\beta_j x_{ji}) \quad (1)$$

where  $h_{j,0}(t)$  is the baseline hazard function and  $\beta_j$  is a parameter. Gene filtering is based on a test of hypothesis  $\beta_j = 0$  (e.g., a score or Wald test [21]). Genes are typically ranked on the basis of the value of absolute standardized test statistic. Gene filtering can be based on the number of genes [4] or a *P*-value cut-off [1,2,5]. A standardized score or Wald test statistic for testing hypothesis  $\beta_j = 0$  is asymptotically normal with unit variance and mean equal to  $D^{1/2} \beta_j \sigma_j^{-1}$  where  $\sigma_j^2$  is the variance of expression levels across patients for gene *j* and *D* is the expected number of events [22]. The gene filtering is thus based on the hazard ratio associated with a change of standard deviation in gene expression for a given number of events.

**Prediction model**

For the set of *K* selected genes ( $j_1, \dots, j_K$ ), the compound covariate for patient *i* is defined as

$$c_i = \sum_{k=1}^K z_{j_k} x_{j_k,i} \quad (2)$$

where  $z_{j_k}$  is the standardized test statistic obtained in the gene filtering for the selected gene  $j_k$  ( $k = 1, \dots, K$ ). The definition of the compound covariates weights by means of standardized test statistics has been suggested for generalized linear models in Radmacher et al. [14]. This weighting policy reflects the criterion in the gene filtering step. Another possible policy is to use an estimate of  $\beta_j$  in stead of  $z_j$  as the weight for gene *j* (e.g., Beer et al. [4]). Our weighting policy gives higher weight to genes with larger variance, which would yield a more robust predictor for subsequent validation studies because the expression profiles for genes with larger variance would be more reproducible.

The compound covariate can be regarded as a prognostic index; patients with large values of the compound covariate may have poor prognosis. We assume the following Cox model to relate the compound covariate to the survival time,

$$h_i(t) = h_0(t) \exp(\psi c_i) \quad (3)$$

## KB en migratie

---

### *Eerste Pagina*

Naam journal: links uitgelijnd

Uitgeverslogo: rechts uitgelijnd

Scheidingslijn: verdelingslijn over de breedte van het blad

Subtitel artikel: links uitgelijnd, lettertype GillSans-Light, lettergrootte 14 pt, font kleur zwart

Open access afbeelding: rechts uitgelijnd, kleur afbeelding

Titel artikel: links uitgelijnd, lettertype GillSans-Bold, lettergrootte 16 pt, font kleur zwart

Auteursnaam: links uitgelijnd, lettertype Giovanni-Book, lettergrootte 15 pt, font kleur zwart

Etc.

### Context

Titel: Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays: methodology article  
Auteur: Shigeyuki Matsui  
Verschenen in: BMC Bioinformatics  
Jaar: 2006

Dit document betreft een wetenschappelijk artikel.

### Gedrag

1. Zoeken in het document.
2. Navigatie. Onderverdeling in Bladwijzers (bladwijzers naar paragrafen in het artikel) en Pagina's (9) waartussen genavigeerd kan worden.
3. Printen
4. Document wijzigen
5. Document samenstellen
6. Kopiëren of extractie van inhoud (tekst)
7. Inhoud uitnemen voor toegang
8. Opmerkingen
9. Invullen van form velden
10. Handtekening
11. Sjabloonpagina's maken

NB: Waar eindigen de eigenschappen van het document en beginnen de eigenschappen van de applicatie waarmee gekeken wordt?

In dit voorbeeld zijn de eigenschappen van het document bij creatie van het document zelf "aan" of "uit" te zetten. Scrollen wordt bijvoorbeeld gezien als eigenschap van de applicatie waarmee het voorbeelddocument bekeken kan worden en niet van het document zelf.

### 1.2 Voorbeeld technische karakterisering van een e-Depot artikel

Technische karakterisering kan voorkomen op bestandsformaatniveau of op bestandsniveau. Karakterisering op bestandsniveau zou kunnen bestaan uit een overzicht van technische metadata die het programma Jhove maakt. Informatie op het niveau van het bestandsformaat zou bijvoorbeeld de aanwezigheid van een header, waarin informatie is opgeslagen over auteur of beschrijving, kunnen zijn.

Jhove output (delen ervan) bij dit voorbeeld artikel:

```
Jhove (Rel. 1.0, 2005-05-26)
Date: 2006-08-16 12:15:42 CEST
RepresentationInformation: test_migratie\t1_wordtopdf\1471-2105-7-156.pdf
ReportingModule: PDF-hul, Rel. 1.4 (2005-03-09)
LastModified: 2006-08-15 23:15:52 CEST
Size: 381971
Format: PDF
Version: 1.3
Status: Not well-formed
SignatureMatches:
  PDF-hul
ErrorMessage: Invalid destination object
Offset: 370422
MIMEtype: application/pdf
PDFMetadata:
  Objects: 477
  FreeObjects: 1
  IncrementalUpdates: 2
  DocumentCatalog:
    PageLayout: SinglePage
    PageMode: UseOutlines
  Outlines:
    Item:
      Title: Abstract
      Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1ce2dd4
    Children:
      Item:
        Title: Background
        Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@122cdb6
      Item:
        Title: Results
        Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1ef9157
      Item:
        Title: Conclusion
        Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@12f0999
    Item:
      Title: Background
      Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@11f2ee1
```

## KB en migratie

---

Item:

Title: Results

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@3ecfff

Children:

Item:

Title: Gene filtering

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@65a77f

Item:

Title: Prediction model

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1d7ad1c

Item:

Title: Predictive accuracy

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@a61164

Item:

Title: Adjustment for prognostic factors

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@bfc8e0

Item:

Title: Simulated data

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@11d0a4f

Item:

Title: Lymphoma data

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@18fd984

Item:

Title: Discussion

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@111a775

Item:

Title: Conclusion

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@91cee

Item:

Title: Acknowledgements

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@4a63d8

Item:

Title: References

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1e0ff2f

Info:

Title: 1471-2105-7-156.fm

Author: csproduction

Creator: FrameMaker 7.0

Producer: Acrobat Distiller 5.0.5 (Windows)

CreationDate: Tue Aug 15 14:03:26 CEST 2006

ModDate: Tue Aug 15 14:10:44 CEST 2006

ID: 0x5c55fb76e1246f68adb2f0ee3c6e88d4, 0x7ef8d9594b611e50de0824ea629f539a

Filters:

FilterPipeline: FlateDecode

Images:

Image:

NisoImageMetadata:

MIMEType: application/pdf

CompressionScheme: Deflate

## KB en migratie

---

ImageWidth: 706

ImageLength: 706

Fonts:

Type0:

Font:

BaseFont: KNLBEF+MT-Extra

Encoding: Identity-H

ToUnicode: true

Font:

BaseFont: KNLANH+SymbolMT

Encoding: Identity-H

ToUnicode: true

Type1:

Font:

BaseFont: Courier

FirstChar: 32

LastChar: 32

FontDescriptor:

FontName: Courier

Flags: FixedPitch, Serif, Nonsymbolic

FontBBox: -28, -250, 628, 805

Encoding: WinAnsiEncoding

Font:

BaseFont: KNLOJI+CXGUNB+Times-Italic

FontSubset: true

FirstChar: 80

LastChar: 80

FontDescriptor:

FontName: KNLOJI+CXGUNB+Times-Italic

Flags: Nonsymbolic, Italic

FontBBox: 0, 0, 605, 653

FontFile3: true

Encoding: WinAnsiEncoding

```
XMP: <rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:iX='http://ns.adobe.com/iX/1.0/'><rdf:Description about="
xmlns='http://ns.adobe.com/pdf/1.3/' xmlns:pdf='http://ns.adobe.com/pdf/1.3/'
pdf:CreationDate='2006-08-15T12:03:26Z' pdf:ModDate='2006-08-15T12:10:44Z'
pdf:Producer='Acrobat Distiller 5.0.5 (Windows)' pdf:Author='csproduction'
pdf:Creator='FrameMaker 7.0' pdf>Title='1471-2105-7-156.fm'/>
<rdf:Description about=" xmlns='http://ns.adobe.com/xap/1.0/'
xmlns:xap='http://ns.adobe.com/xap/1.0/' xap:CreateDate='2006-08-15T12:03:26Z'
xap:ModifyDate='2006-08-15T12:10:44Z' xap:Author='csproduction'
xap:MetadataDate='2006-08-15T12:10:44Z'><xap:Title><rdf:Alt><rdf:li xml:lang='x-
default'>1471-2105-7-156.fm</rdf:li></rdf:Alt></xap:Title></rdf:Description>
<rdf:Description about=" xmlns='http://purl.org/dc/elements/1.1/'
xmlns:dc='http://purl.org/dc/elements/1.1/' dc:creator='csproduction' dc:title='1471-2105-7-
156.fm'/>
</rdf:RDF>
```

## KB en migratie

---

### 1.3 Overzicht (voorlopige) functionele karakterisering van inhoud e-Depot

Hieronder een eerste opzet voor het indelen van de inhoud van het e-Depot op karakterisering van de opgeslagen objecten. In het schema wordt aangegeven welke van de vijf karakteristieken van het digitale object minimaal behouden moeten blijven bij eventuele migraties.<sup>1</sup>

Type object	Functioneel doel van het object			
	Publicatie	Educatief programma	Digitale master cultureel erfgoed	
Tekstdocument	Inhoud Structuur			
Afbeelding			Inhoud Structuur Uiterlijk	
Spreadsheet				
Interactieve software		Inhoud Structuur Uiterlijk Gedrag Context?		
Website	Inhoud Structuur Uiterlijk? Gedrag?			

---

<sup>1</sup> Het overzicht is een eerste opzet voor een indeling. Om de uitgangspunten van het e-Depot aan te vullen, zou een officiële indeling van de inhoud van het e-Depot gemaakt moeten worden, met breed draagvlak binnen de afdelingen van de KB.

# KB en migratie

---

	Context?			
--	----------	--	--	--